CLAIMS

I claim:

1.   A computer-implemented method comprising:

defining a set of reduced regular expressions for particular patterns in strings; and

learning, from a training set, a knowledge base that uses the reduced regular expressions to resolve ambiguity based upon the strings in which the ambiguity occurs, wherein the learning includes transformation sequence learning to create a set of rules that use the reduced regular expressions to resolve ambiguity based upon the strings in which the ambiguity occurs.

2.   A computer-implemented method as recited in claim 1, wherein the set of reduced regular expressions are defined over a finite alphabet Σ, wherein the alphabet is a union of multiple sets of distinct classes.

3.   A computer-implemented method as recited in claim 1, wherein the training set comprises a labeled corpus.

4.   A computer-implemented method as recited in claim 1, wherein the set of reduced regular expressions specify types of patterns that are allowed to be explored when learning from the training set.

5. A computer-implemented method as recited in claim 1, wherein the learning includes applying a set of very reduced regular expressions that are a proper subset of the reduced regular expressions.

6. A computer readable medium having computer-executable instructions that, when executed on a processor, perform a method comprising:

defining a set of reduced regular expressions for particular patterns in strings; and

learning, from a training set, a knowledge base that uses the reduced regular expressions to resolve ambiguity based upon the strings in which the ambiguity occurs, wherein the set of reduced regular expressions specify types of patterns that are allowed to be explored when learning from the training set.

7. A computer readable medium as recited in claim 6, wherein the set of reduced regular expressions are defined over a finite alphabet Σ, wherein the alphabet is a union of multiple sets of distinct classes.

8. A computer-implemented method as recited in claim 6, wherein the training set comprises a labeled corpus.

9. A computer-implemented method as recited in claim 6, wherein the learning comprises transformation sequence learning to create a set of rules that use the reduced regular expressions to resolve ambiguity based upon the strings in which the ambiguity occurs.

10.   A computer-implemented method as recited in claim 6, wherein the learning includes applying a set of very reduced regular expressions that are a proper subset of the reduced regular expressions.

11.   A computer-implemented method comprising:

receiving a string with an ambiguity site;

applying reduced regular expressions to describe a pattern in the string, wherein the reduced regular expressions:

are included in a knowledge base that is learned from a training set; and

specify types of patterns that are allowed to be explored when the knowledge base is learned; and

selecting one of the reduced regular expressions to resolve the ambiguity site.

12.   A computer-implemented method as recited in claim 11, wherein the applying includes applying a set of very reduced regular expressions that are a proper subset of the reduced regular expressions.

13.   A computer-implemented method comprising:

receiving a string with an ambiguity site;

applying reduced regular expressions to describe a pattern in the string, wherein the applying includes applying a set of very reduced regular expressions that are a proper subset of the reduced regular expressions; and

selecting one of the reduced regular expressions to resolve the ambiguity site.

14. A computer readable medium having computer-executable instructions that, when executed on a processor, perform a method comprising:

receiving a string with an ambiguity site;

applying reduced regular expressions to describe a pattern in the string, wherein:

the reduced regular expressions are included in a knowledge base that is learned from a training set; and

the reduced regular expressions specify types of patterns that are allowed to be explored when the knowledge base is learned; and

selecting one of the reduced regular expressions to resolve the ambiguity site.

15. A computer readable medium as recited in claim 14, wherein the applying includes applying a set of very reduced regular expressions that are a proper subset of the reduced regular expressions.

16. A computer readable medium having computer-executable instructions that, when executed, direct a computer to:

read a training set;

construct a graph having a root node that contains a primary position set of the training set and multiple paths from the root node to secondary nodes that

represents a reduced regular expression, the secondary node containing a secondary position set to which the reduced regular expression maps;

score the secondary nodes to identify a particular secondary node; and

identify the reduced regular expression that maps the path from the root node to the particular secondary node.

17.   A training system comprising:

a memory to store a training set;

a processing unit; and

a disambiguation trainer, executable on the processing unit, to define a set of reduced regular expressions for particular patterns in strings of the training set and learn a knowledge base that uses the reduced regular expressions to describe the strings wherein the reduced regular expressions specify types of patterns that are allowed to be explored when the knowledge base is learned from the training set.

18.   A training system as recited in claim 17, wherein the training set comprises a labeled corpus.

19.   A training system as recited in claim 17, wherein the disambiguator trainer employs transformation sequence learning to create a set of rules that use the reduced regular expressions to describe the strings.

20.   A system comprising:

a memory to store a knowledge base that uses reduced regular expressions to resolve ambiguity based upon strings in which the ambiguity occurs, wherein the knowledge base is learned from a training set using the reduced regular expressions, the reduced regular expressions specify types of patterns that are allowed to be explored when the knowledge base is learned;

a processing unit; and

a disambiguator, executable on the processing unit, to receive a string with an ambiguity site and apply a reduced regular expression from the knowledge base that describes a pattern in the string to resolve the ambiguity site.